Use of Adjoint Physics for 4D VAR with the NCEP Global Spectral Model

Presentation for Ph.D. Dissertation

by

Shaoqing Zhang

Department of Meteorology

The Florida State University

What is variational analysis?

What is 4D VAR? What good is it?

- Variational analysis: vary control parameters to adjust system to optimal state.

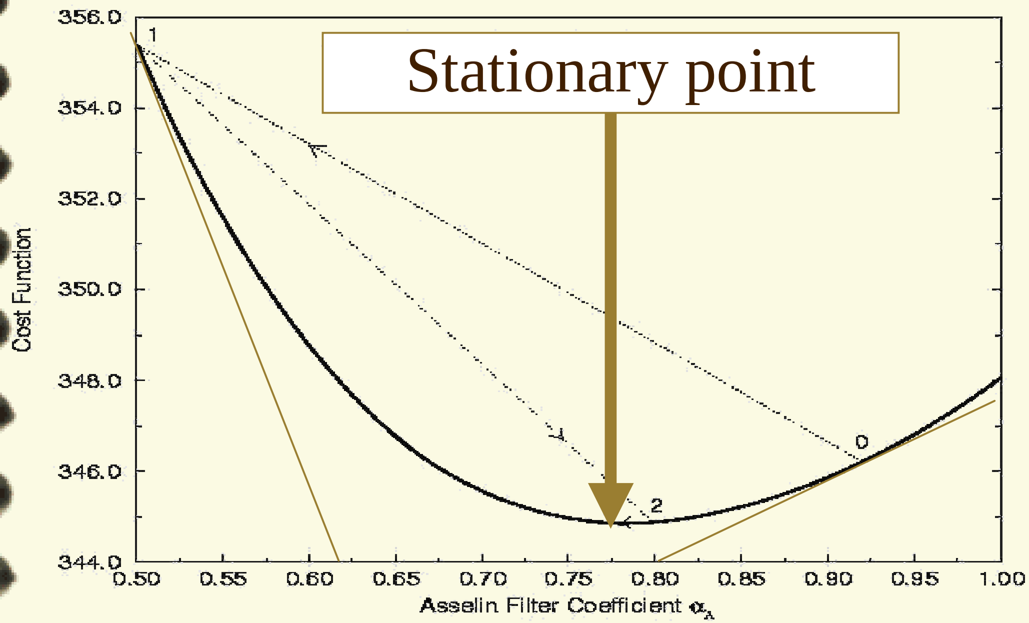
  Control system: NWP model

  Control parameters: ICs, physical parameters

  Optimal state: Minimal forecast errors (cost function).

- 4D VAR: apply variational analysis to minimize a defined cost function over space and time.
- Application: Find an optimal estimate of ICs or parameters, which is internally consistent between model dynamics and observations.

What are minimization and adjoint?



- ✓ A minimization algorithm seeks a stationary point: evaluate cost function and its gradient
- ✓ An adjoint integration efficiently evaluates cost function gradient.

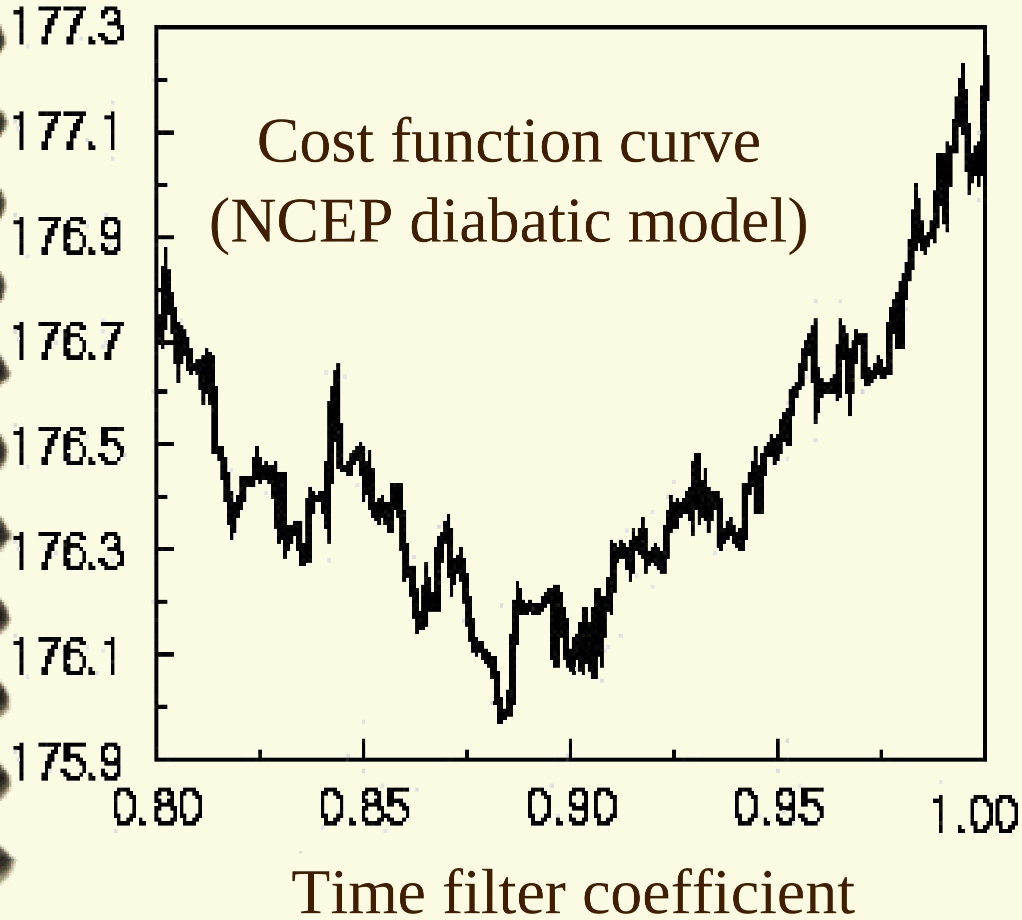
Adjoint sensitivity analysis

ICs & parameters

Error (cost ftn)

Nonlinear Forecast

Challenge: Discontinuous physics



- ✓ Discontinuous physics ⇒ discontinuous cost function.
- ✓ Past approach: smooth discontinuities in physics.
- ✓ Smoothing introduces many additional local minima.

Goals of my research

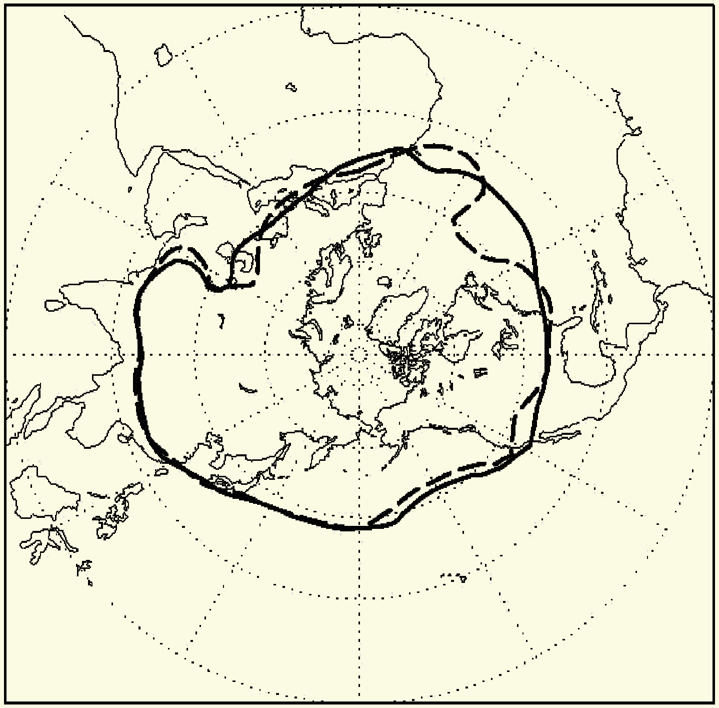
- ✓ Answer questions:
  - – Can an adjoint correctly evaluate the cost function gradient when model physics are discontinuous?
  - – Can a minimization algorithm designed for differentiable functions work for a discontinuous cost function? Do we have a better solution?
- ✓ Construct a variational analysis system with the NCEP global spectral diabatic model
- ✓ Carry out experiments on data assimilation and parameter fitting by 4D VAR approach.

Outline

- ✓ Review of classical approach to variational analysis
  - Lagrange multiplier solves a constrained problem (Sasaki 1958)
  - Optimal control theory (LeDimet & Talagrand 1986)
  - Perturbation analysis (PA) approach to derive the gradient of the cost function (J)
  - Newton and quasi-Newton minimization algorithms.
- ✓ Answers for problems introduced by physics
  - New insight (rather than using PA) between adjoint and gradient:

    Adjoint of discon. physics does work for deriving the gradient of J.
  - Limited Memory Quasi-Newton method (L-BFGS) usually works for minimizing J on discon. physics but sometimes has problems.
  - Bundle method for discontinuous functions is better but slow.
  - Optimal ICs and parameters improve forecasts just for 3 days.
- ✓ Future work

Optimal control problem

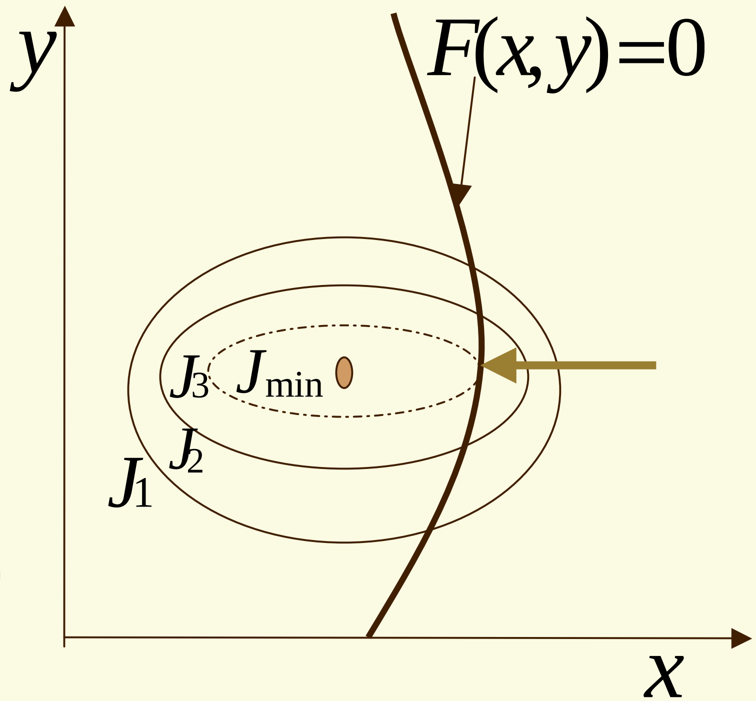


Example: Forecast (solid), observation (dashed)

- Let **x** =column vector of all model variables

  Let **β** =column vector of model parameters
- Let $\frac{\partial \mathbf{x}}{\partial \mathbf{t}} = \mathbf{F}(\mathbf{x}, \boldsymbol{\beta})$ =NWP model
- ***F(x ,β) is discontinuous when parameterized physics are included***
- Let J(**x**) = specified error measurement in a time window (cost function)
- Problem: Find **x** at t=0 and **β** that minimize J

Sasaki (1958): Lagrange multiplier method for constrained problem



- ✓ Lagrange multiplier method constructs a new expression, Lagrangian,

$$L = J(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{F}(\mathbf{x})$$

- ✓ Seek the stationary point $(\mathbf{x}, \boldsymbol{\lambda})$ of the Lagrangian by solving Euler-Lagrange equations

$$\begin{cases} \frac{\partial L}{\partial \mathbf{x}} = 0 \\ \frac{\partial L}{\partial \boldsymbol{\lambda}} = 0 \end{cases}$$

But:

- Too many equations
- Poor convergence
- Too expensive computationally

Le Dimet & Talagrand (1986): Adjoint technique to derive gradient by PA

- ✓ Cost function depends on control variable, $\boldsymbol{\alpha}=(\mathbf{x_0},\boldsymbol{\beta})$, with numerical model as bridge:

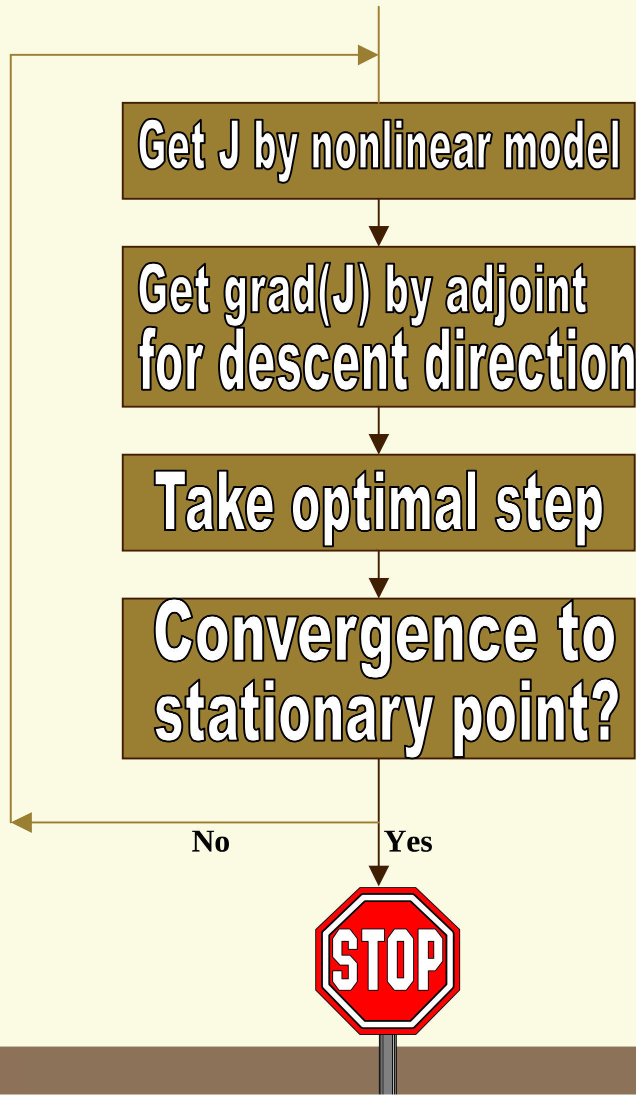
$$\begin{cases} J(\boldsymbol{\alpha}) = \frac{1}{2} < \mathbf{W}(\mathbf{x}-\mathbf{x}^o), (\mathbf{x}-\mathbf{x}^o) > \\ \mathbf{x} = \mathbf{H}\ (\ \boldsymbol{\alpha}\ ) \rightarrow \delta\mathbf{x} = \mathbf{L}\delta\boldsymbol{\alpha} \end{cases} \rightarrow \begin{cases} \delta J = < \mathbf{W}(\mathbf{x}-\mathbf{x}^o), \delta\mathbf{x} > \\ \delta J \overset{\downarrow}{=} < \mathbf{W}(\mathbf{x}-\mathbf{x}^o), \mathbf{L}\delta\boldsymbol{\alpha} > \\ \delta J \overset{\downarrow}{=} < \mathbf{L}^*\mathbf{W}(\mathbf{x}-\mathbf{x}^o), \delta\boldsymbol{\alpha} > \end{cases}$$

- ✓ Since $\delta J = <\nabla|_{\alpha} J, \delta\, \boldsymbol{\alpha}>$, $\nabla|_{\alpha} J = L^* W (\mathbf{x} - \mathbf{x}_0)$, where

  L^* = adjoint of matrix L

- ✓ With the gradient, a minimization algorithm (popularly L-BFGS) can iterate to solve for optimal values of control variables

Newton and quasi-Newton minimization algorithms



- Solve Newton zero roots as an optimal step size (Newton method)

$$J(\boldsymbol{\alpha})=J(\boldsymbol{\alpha}_0)+\nabla_{\alpha} J|_{\alpha_0}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)+\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\mathbf{A}\,(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)$$

$$\downarrow$$

$$\nabla_{\alpha} J|_{\alpha}=\nabla_{\alpha} J|_{\alpha_0}+\mathbf{A}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)$$

$$\downarrow$$

$$\nabla_{\alpha} J|_{\alpha_0}+\mathbf{A}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)=0$$

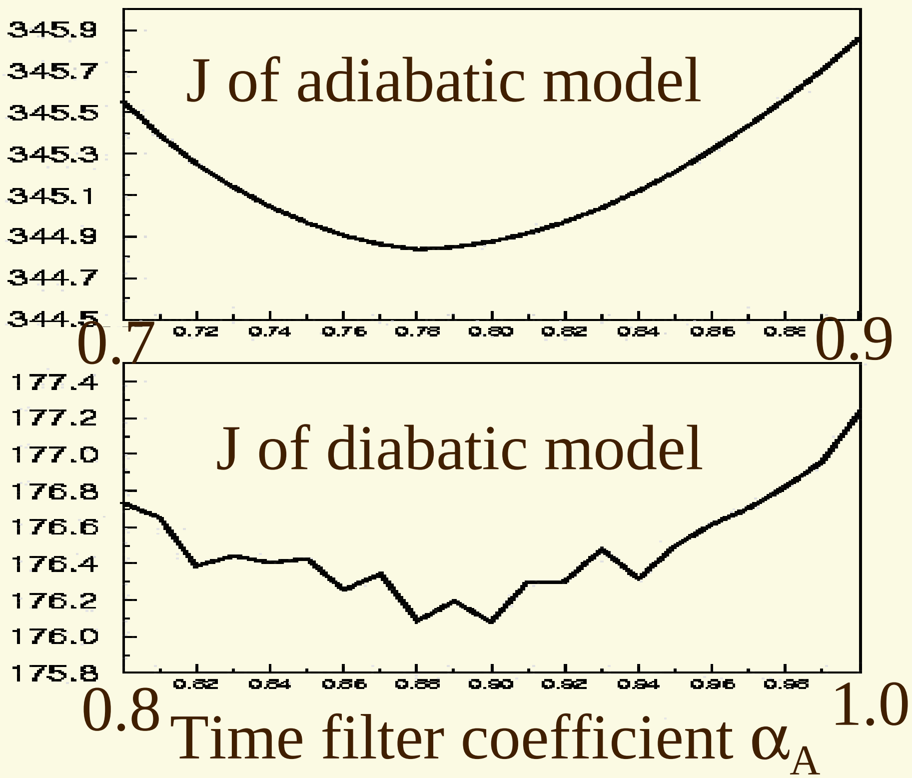
- Iteratively refine the approximation for the inverse of Hessian matrix (quasi-Newton method, L-BFGS)

$$\begin{cases} \mathbf{H}_{n+1}=\mathbf{H}_n+\textbf{correction} \\ \boldsymbol{\alpha}_{n+1}-\boldsymbol{\alpha}_n=\mathbf{H}_{n+1}(\nabla J_{n+1}-\nabla J_n) \\ \lim\limits_{n\to\infty}\mathbf{H}_n=\mathbf{A}^{-1} \end{cases}$$

- Line search to determine optimal stepsize $\beta^{(k)}$

Steps to develop adjoint

- ✓ Code and test TLM: compare $\mathbf{x}(\alpha+\delta\alpha)-\mathbf{x}(\alpha)$ and $\mathbf{L}\ \delta\alpha$
- ✓ Code and test adjoint: compare $J(\alpha+\delta\alpha)-J(\alpha)$ and $\mathrm{grad}(J)\ \delta\alpha$
- ✓ TLM and adjoint tests by PA fail with discont. physics



Ex: Asselin Filter:

$$\tilde{A}(t)=\tilde{A}(t-1)+\varepsilon[\tilde{A}(t-1)-2A(t)+A(t+1)]$$

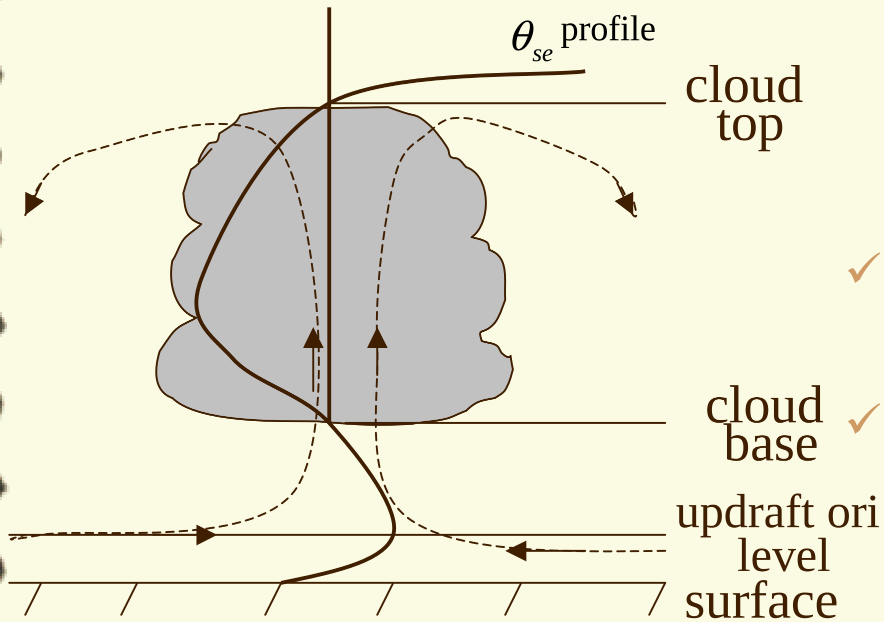
$$\begin{cases} \alpha_A = 1-2\varepsilon \\ \alpha_A = \alpha_{A0} + \Delta\alpha_A \times n \\ \Delta\alpha_A = 0.01 \end{cases}$$

TLM test and gradient test of adjoint based on perturbation analysis

- ✓ Compute $\{\mathbf{x}(\alpha+\beta\delta\alpha) - \mathbf{x}(\alpha)\} / \mathbf{L}\,\beta\delta\alpha$
- ✓ Compute $\{J(\alpha+\beta\delta\alpha) - J(\alpha)\} / \mathrm{grad}(J)\,\beta\delta\alpha$

TLM test			Gradient test of adjoint		
$\log\beta$	ADB-model	DB-model	$\log\beta$	ADB - Model	DB - Model
-1	0.99998	4173.94	-8	1.06577	2.24501
-2	0.99999	12.0942	-9	1.00622	0.25756
-3	1.00000	1.00045	-10	1.00063	-4.1876
-4	1.00000	1.00017	-11	1.00006	1.00129
-5	1.00000	—	-12	1.00000	1.00392
-6	1.00000	1.00031	-13	0.99998	1.00076
-7	1.00000	1.00066	-14	1.00076	0.99967
-8	1.00000	—	-15	0.99849	0.98182
-9	1.00000	—	-16	0.98786	0.81740

Example of discontinuities in a simplified Arakawa-Schubert cumulus parameterization



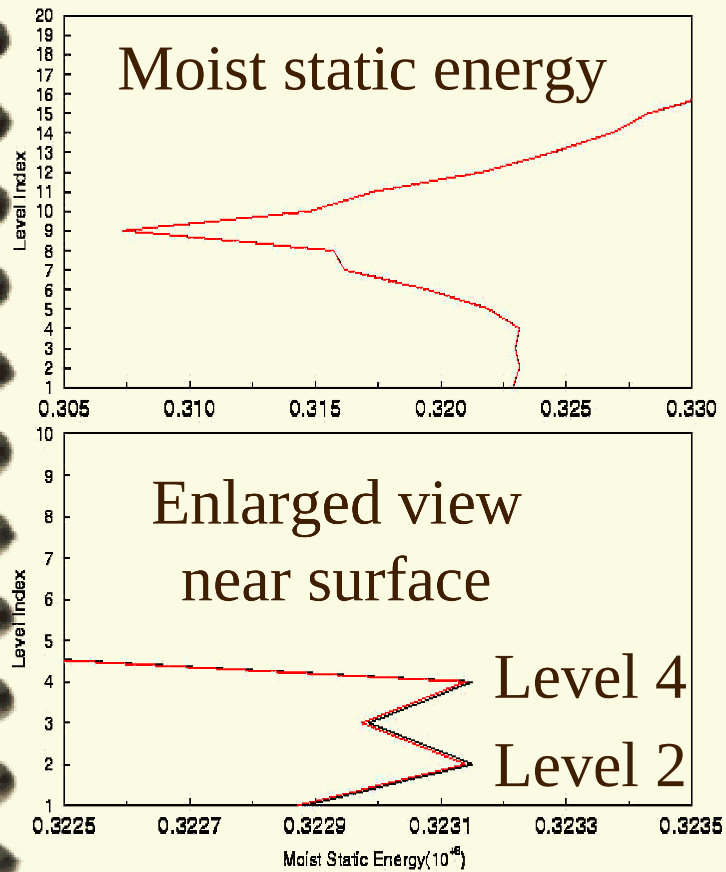
A-S schematic diagram for one cloud type

- Conditional instability defines updraft originating level; cloud base at lifting condensation level; cloud top where parcel θ_{se} equals environment θ_{se}
- 150 hPa is threshold for updraft layer and cloud thickness
- Check diabatic model behavior

  $\mathbf{x}(\alpha+\delta\alpha) - \mathbf{x}(\alpha)$ for initial field on

  1 Nov 1995
- Choose column where initial cumulus is turned off after small change in θ_{se} profile

Example of discontinuities in a simplified Arakawa-Schubert cumulus parameterization

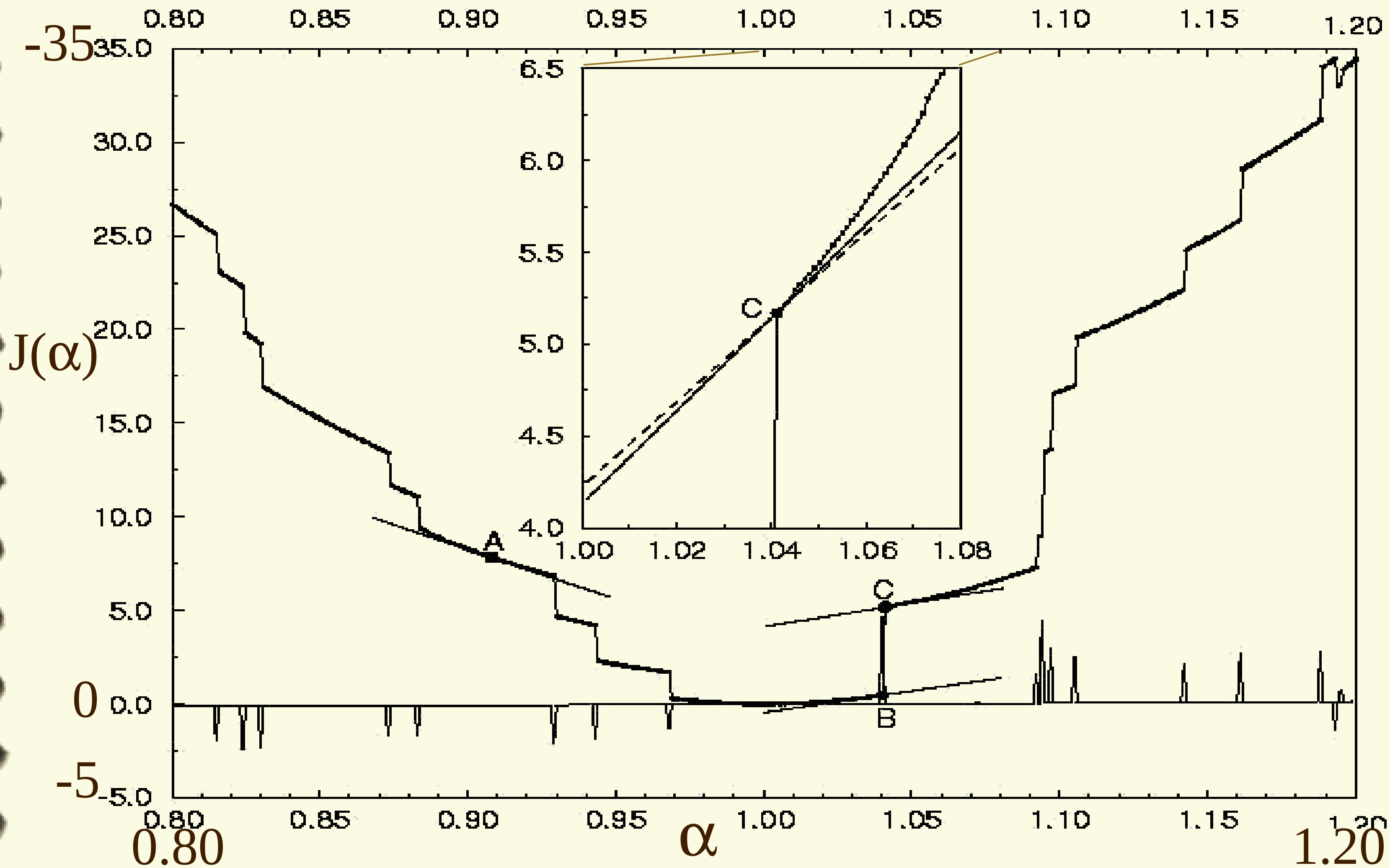


- ✓ Example: column 212, latitude circle 39, time step 3
- ✓ Updraft depth: 154 hPa when level 2 is the updraft originating level; 68 hPa for level 4
- ✓ Any small perturbation may cause cumulus to be turned on/off suddenly
- ✓ Model response jumps

$J(\alpha)=\sum w[f(T,q)-f^{obs}]^2$, f = Arakawa-Sch. parameterization

$T=T_{3.75^oN}+\alpha(T_{1.875^oN}-T_{3.75^oN})$ (28 levels ×384 columns)

by $\alpha=0.8+0.001\times n$ and $f^{obs}=f(T_{1.875^oN},\ q_{1.875^oN})$ on 11/01/95



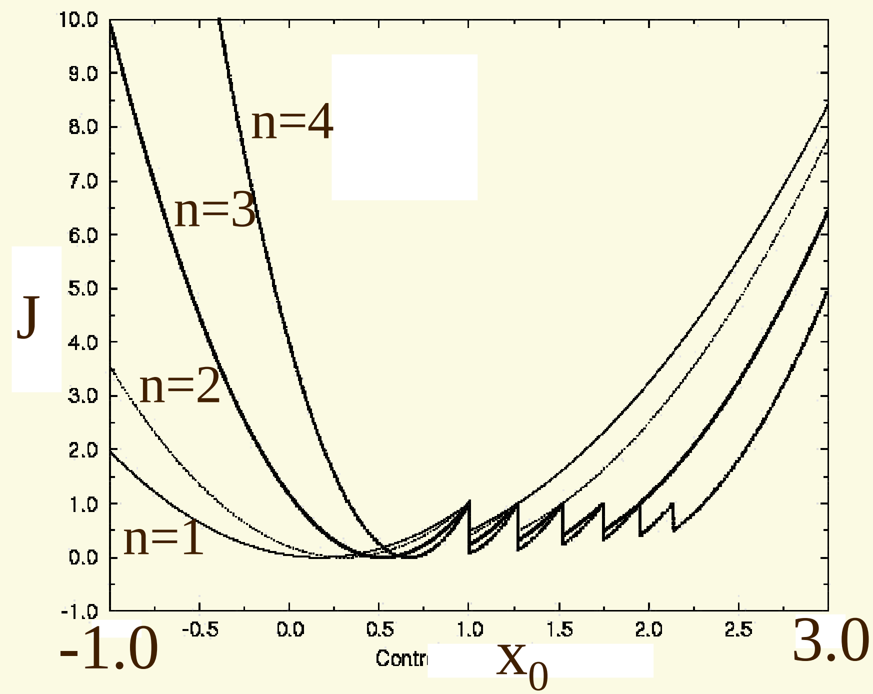
Character of cost function with discontinuous physics: Simple model

- ✓ The cost function, J, is piecewise differentiable due to piecewise differentiable source term (physics).
- ✓ For k thresholds and n time steps, max number of differentiable segments of J is $k \cdot 2^n$

$$\frac{\partial x}{\partial t} = \begin{cases} f_1(x), & x < x_c \\ f_2(x), & x \geq x_c \end{cases}$$

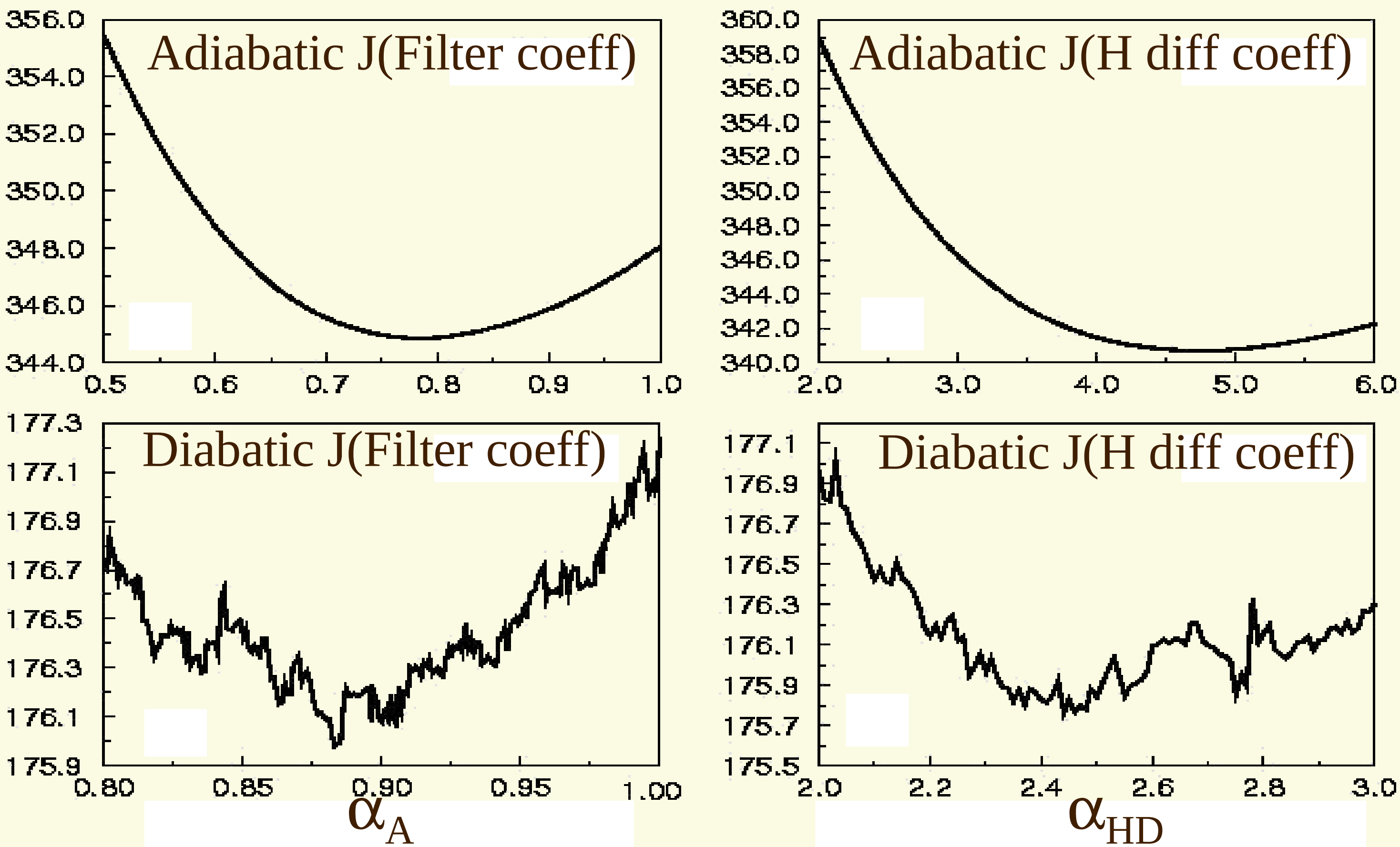
$f_1(x) = 2x-2$, $f_2(x) = x-4$,
$x_c = 1$, $dt = 0.1$.

Cost function:
$J = x^2(t_n)$



Character of cost function with discontinuous physics: Real model

- ✓ A real numerical model is an extension of the single-variable model on grids and variables



Can adjoint of discontinuous physics find cost function gradient? Theory:

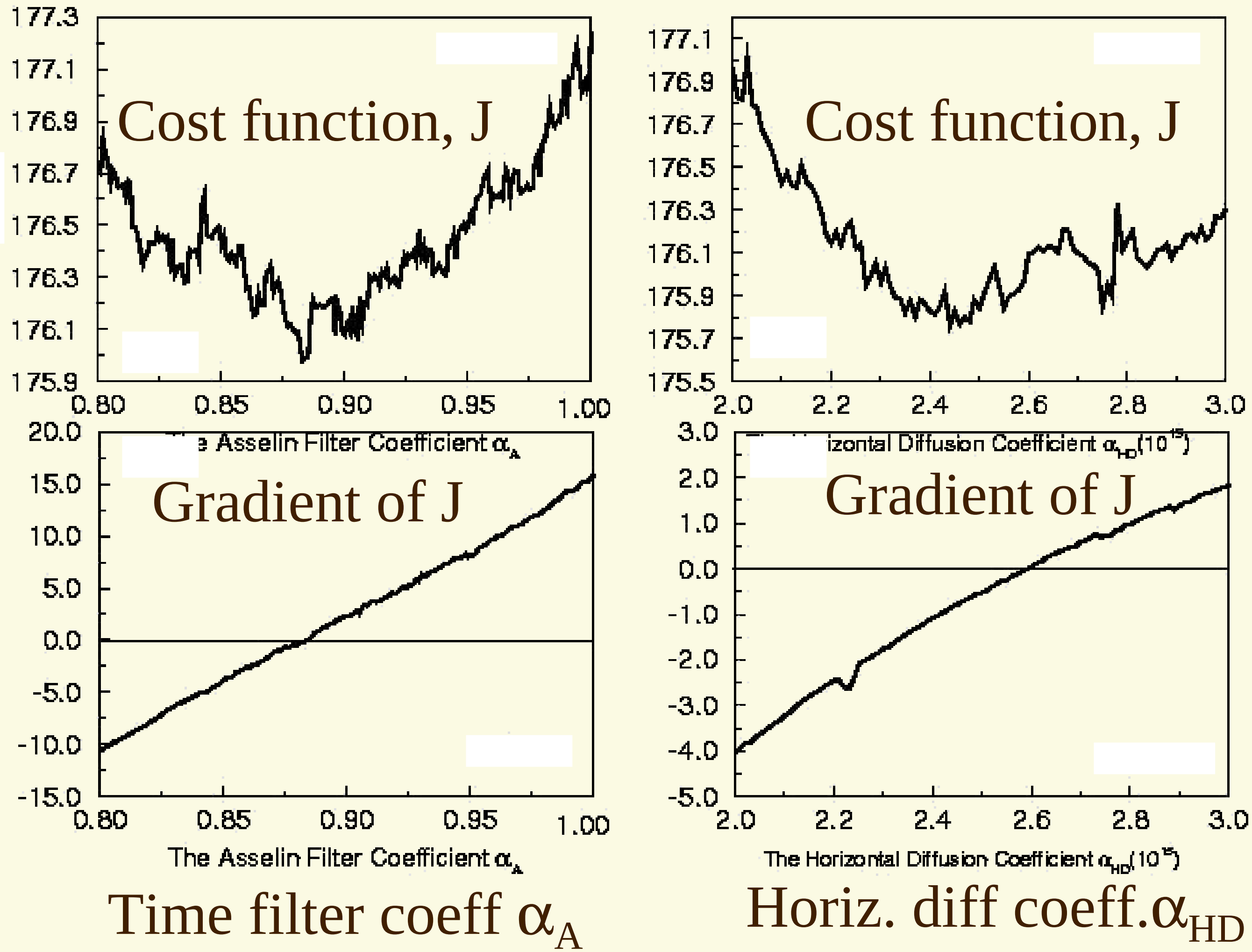
- ✓ Gradient of J of a single-variable model w.r.t. IC is evaluated by chain rule of differentiation, and every integration time step forms a sub-function:

$$\frac{dJ}{dx_0}=\frac{dJ}{dx_n}\frac{dx_n}{dx_{n-1}}\cdots\frac{dx_2}{dx_1}\frac{dx_1}{dx_0}=\left(\frac{dx_1}{dx_0}\frac{dx_2}{dx_1}\cdots\frac{dx_n}{dx_{n-1}}\right)\frac{dJ}{dx_n}\ \text{=adjoint integration of}\ \frac{dJ}{dx_n}$$

- ✓ For multi-variable models, expanding the chain rule forms the integration of an adjoint model:

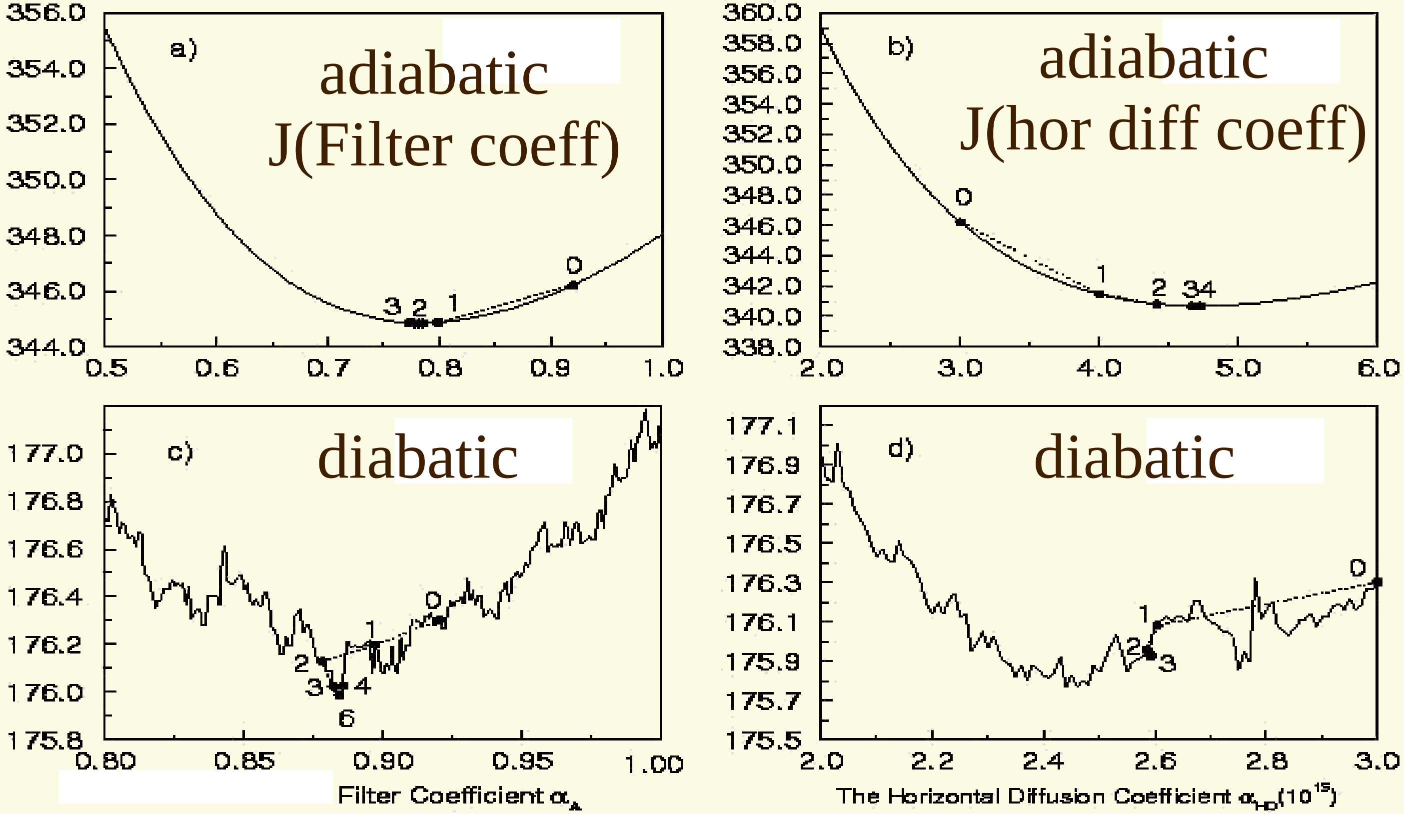
$$\nabla|_{\mathbf{x}_0}J=\begin{pmatrix}\frac{\partial J}{\partial x_{01}}\\ \frac{\partial J}{\partial x_{02}}\\ \vdots\\ \frac{\partial J}{\partial x_{0n}}\end{pmatrix}=\begin{pmatrix}\frac{\partial J}{\partial x_1}\frac{\partial x_1}{\partial x_{01}}+\frac{\partial J}{\partial x_2}\frac{\partial x_2}{\partial x_{01}}+\cdots+\frac{\partial J}{\partial x_n}\frac{\partial x_n}{\partial x_{0n}}\\ \frac{\partial J}{\partial x_1}\frac{\partial x_1}{\partial x_{02}}+\frac{\partial J}{\partial x_2}\frac{\partial x_2}{\partial x_{02}}+\cdots+\frac{\partial J}{\partial x_n}\frac{\partial x_n}{\partial x_{02}}\\ \vdots\\ \frac{\partial J}{\partial x_1}\frac{\partial x_1}{\partial x_{0n}}+\frac{\partial J}{\partial x_2}\frac{\partial x_2}{\partial x_{0n}}+\cdots+\frac{\partial J}{\partial x_n}\frac{\partial x_n}{\partial x_{0n}}\end{pmatrix}=\mathbf{L}_1^T\mathbf{L}_2^T\cdots\mathbf{L}_{t_R}^T\frac{\partial J}{\partial \mathbf{x}}=\mathrm{L}^{\mathrm{T}}\frac{\partial \mathrm{J}}{\partial \mathrm{x}}$$

Can adjoint of discontinuous physics find cost function gradient? Yes.



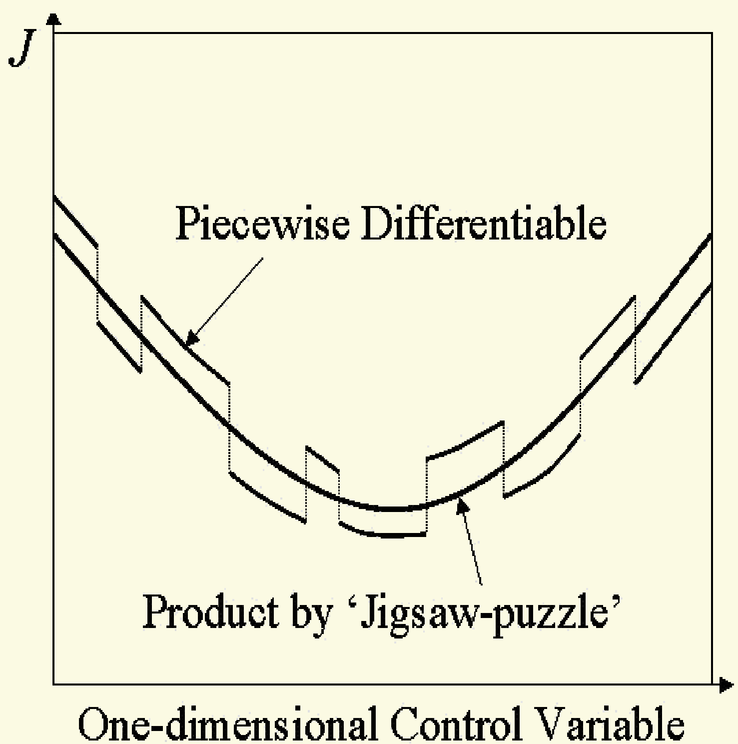
Can quasi-Newton algorithm minimize a piecewise differentiable J? Often yes.

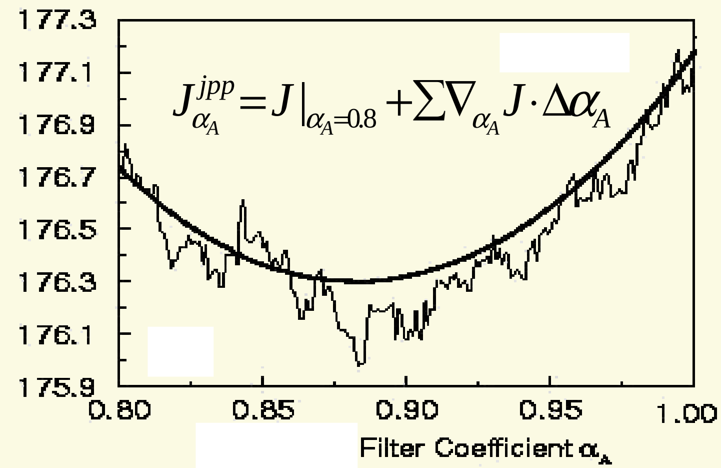
- ✓ L-BFGS algorithm can usually find the stationary point of J with the correct gradient evaluated from the adjoint.
- ✓ The stationary point may not be the global minimum.

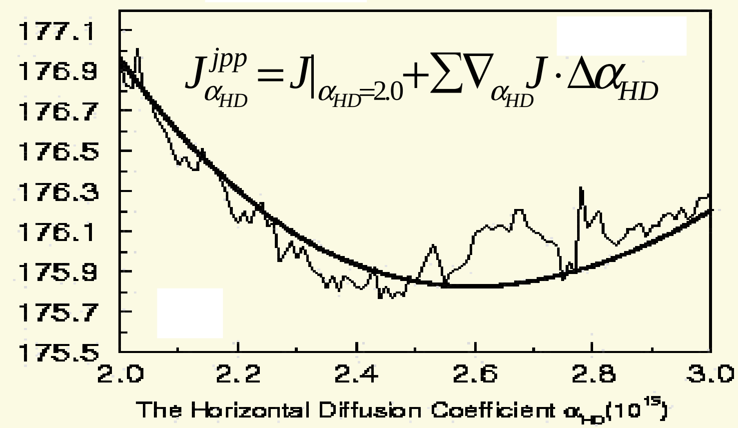


Can quasi-Newton algorithm minimize a piecewise differentiable J? Often yes.

- ✓ Rough curve = actual cost function
- ✓ Smooth curve = integral of gradient. Similar minima.







L-BFGS algorithm (quasi-Newton) minimizes piecewise differentiable J for 282 of 300 ICs.

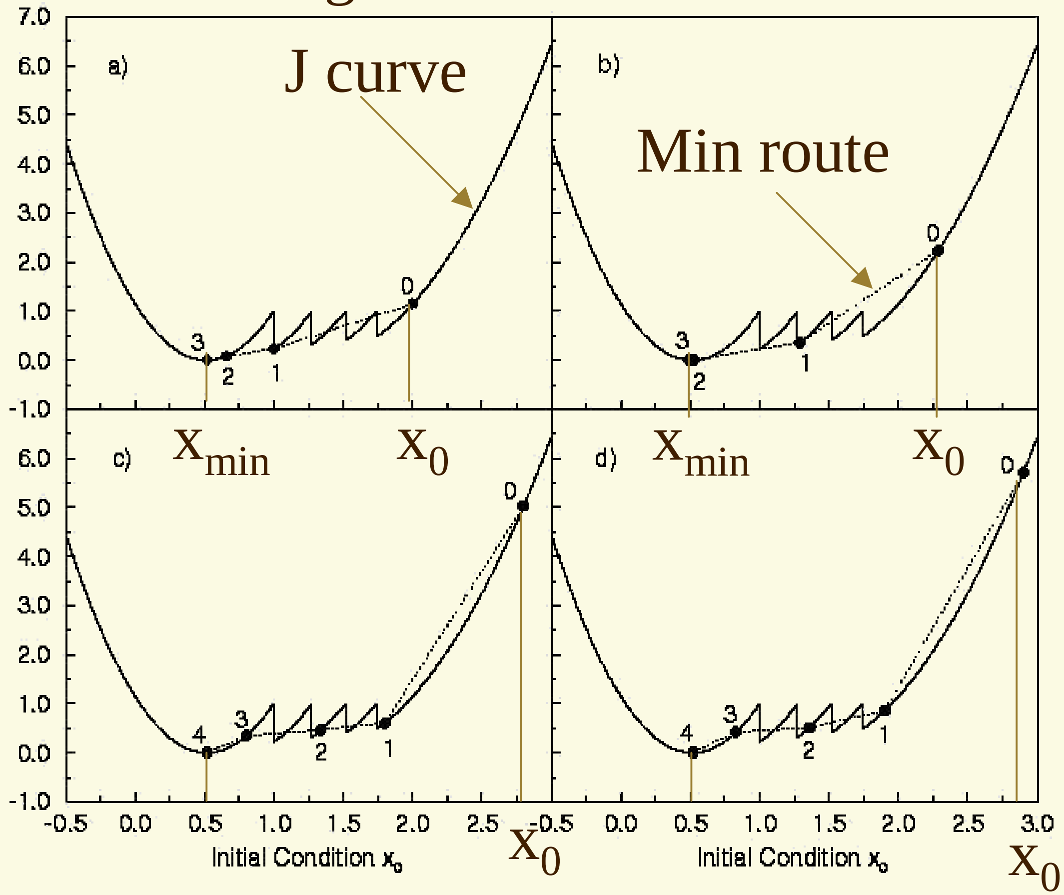
Test case:

$$\frac{\partial x}{\partial t} = \begin{cases} 2x - 2, & x < 1 \\ x - 4, & x \geq 1 \end{cases}$$

Integrate forward
4 steps with dt = 0.1

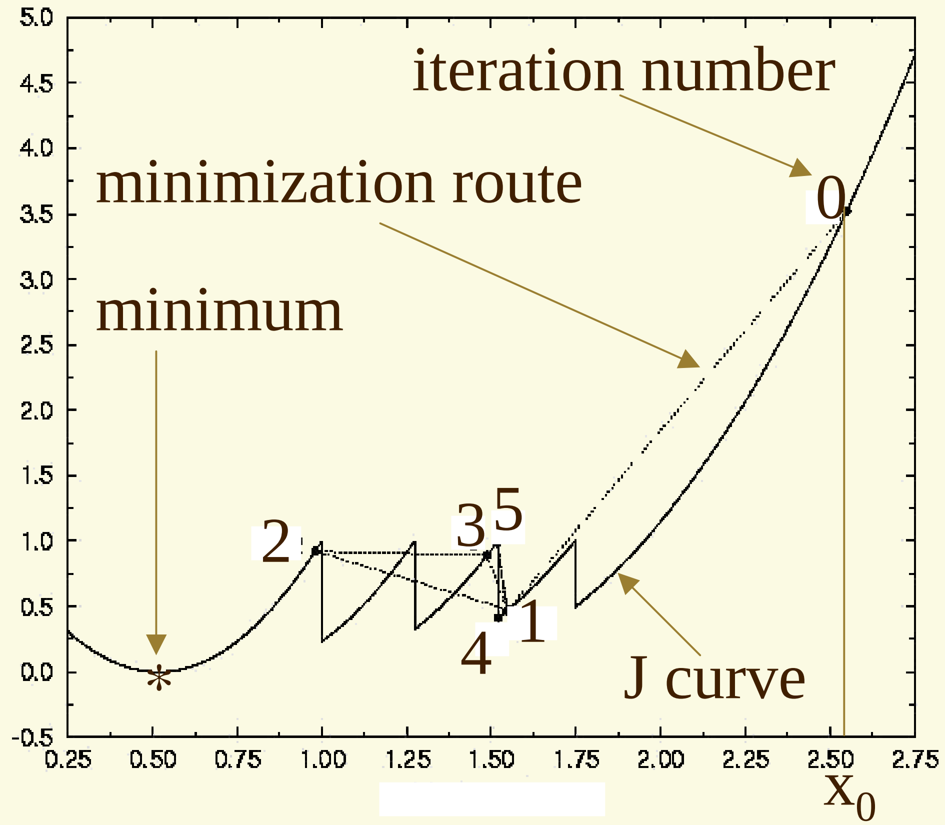
Cost function:
J = x^2 at t_4

Successful minimization of J starting from 4 different ICs



One example from the 18 failed cases: Same equations, different initial guess for x_0

- ✓ Minimization trapped by discontinuity
- ✓ Algorithm works if jump size reduced.

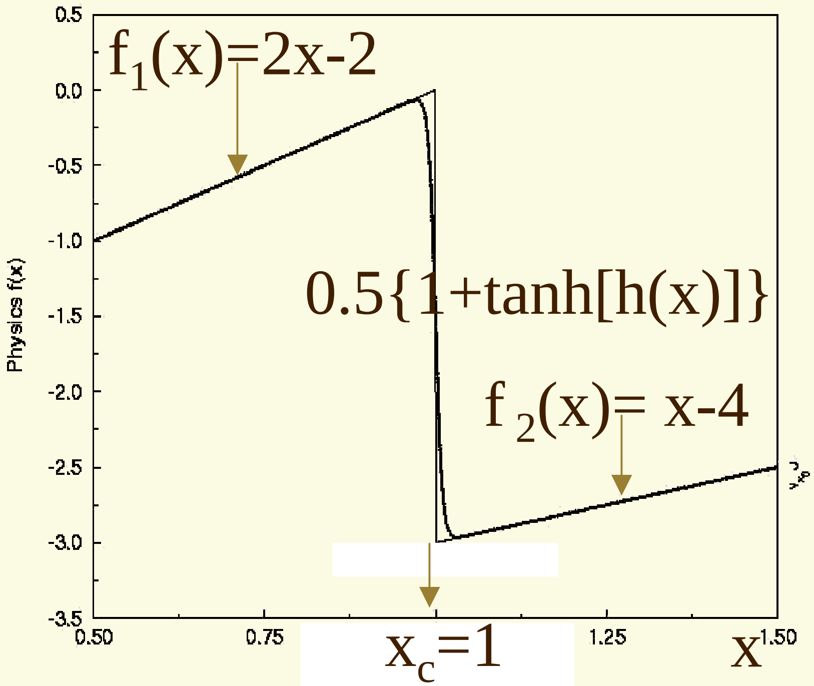


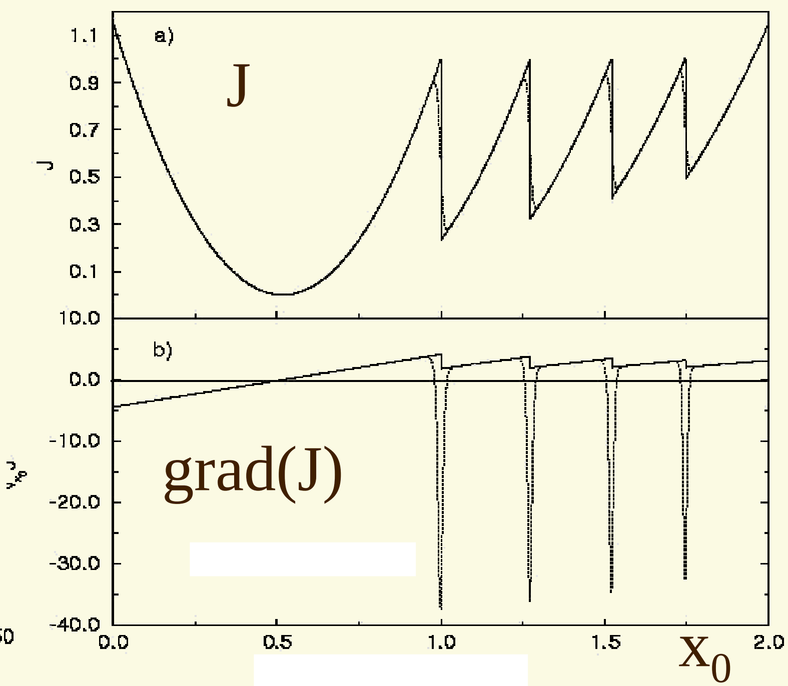
Smoothing the discontinuity

- ✓ Remove discontinuity in a simple model using smooth ftn.

$$\frac{\partial x}{\partial t} = \begin{cases} f_1(x), & x<x_c \\ f_2(x), & x\geq x_c \end{cases}$$

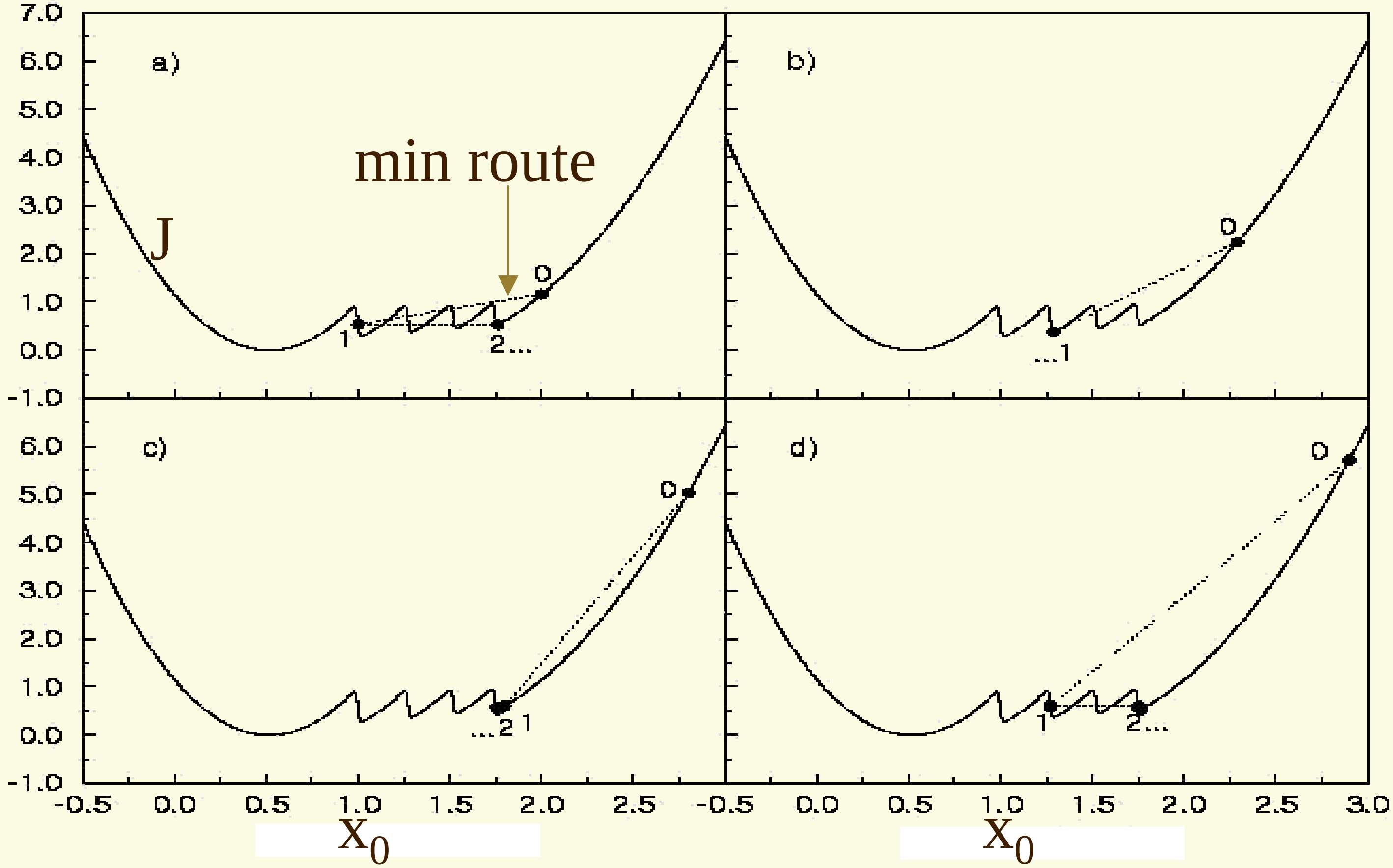
- ✓ Smoothing introduces extra stationary points for J.





Impact of smoothing: L-BFGS finds minimum in only 263 of 300, 37 failures.

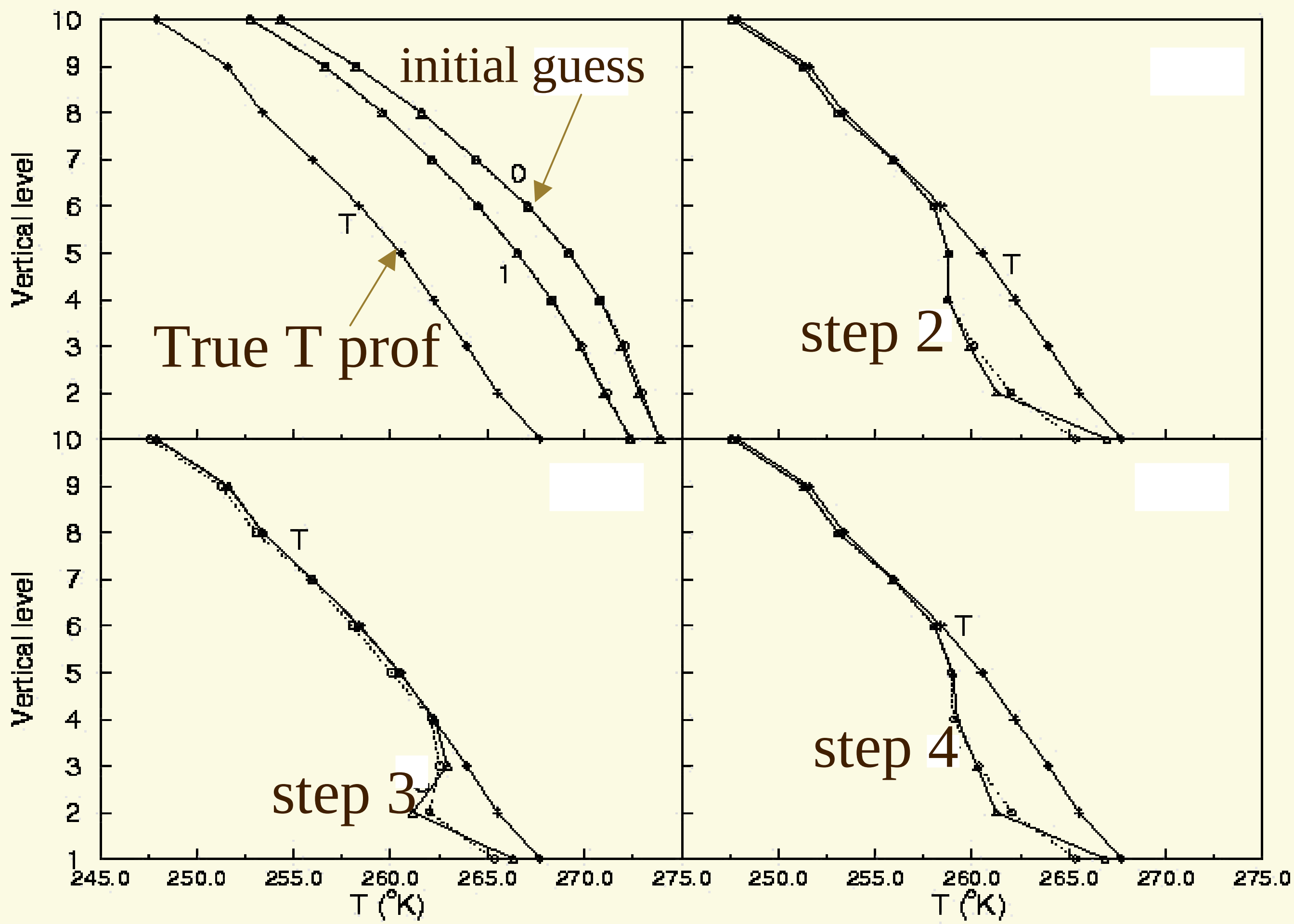
Same 4 ICs as in previous example:
Now trouble.



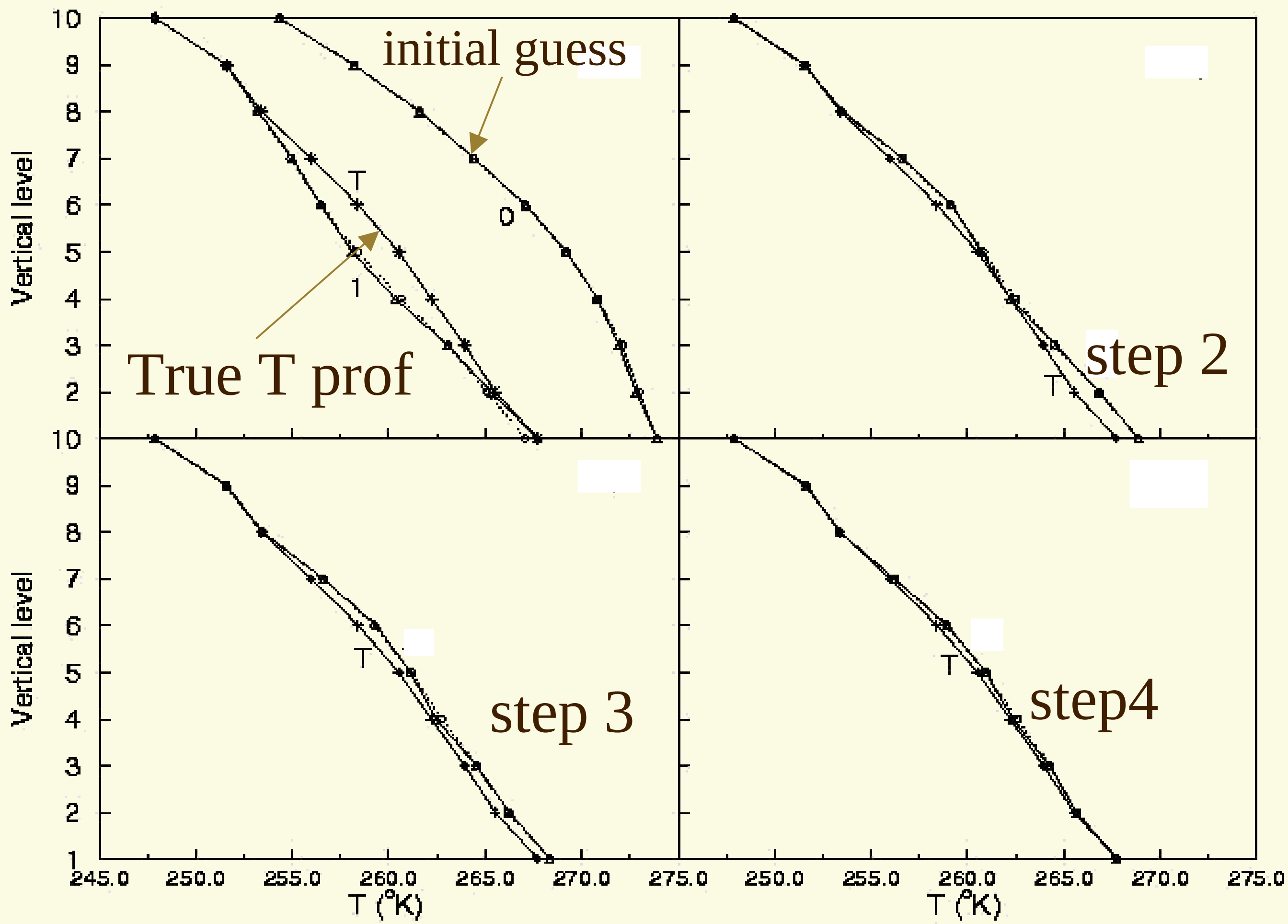
Can we do better? Test piecewise-differentiable "bundle" algorithm

- ✓ Test: $J = \sum w[f(\text{guess}) - f(\text{truth})]^2\}$
- ✓ Take f_i = shallow convection operator
  Discontinuities: conditional instability defines cloud base (lifting condensation level) and cloud top (highest instability level); different diffusion coefficients for different layers.
- ✓ Truth = T & q profiles for column 111 at 12N, 1 July 1995
- ✓ Initial guess = T & q profile for some other column at 12N or 12S. Try each column.
- ✓ Iterate to minimize J (to 0) using L-BFGS or bundle method
- ✓ Bundle method uses a bundle of gradients (side-grad) to construct a sub-gradient to force J to decrease.
- ✓ L-BFGS fails for 3 of 383 columns. Bundle method works for all, but computational cost almost double.

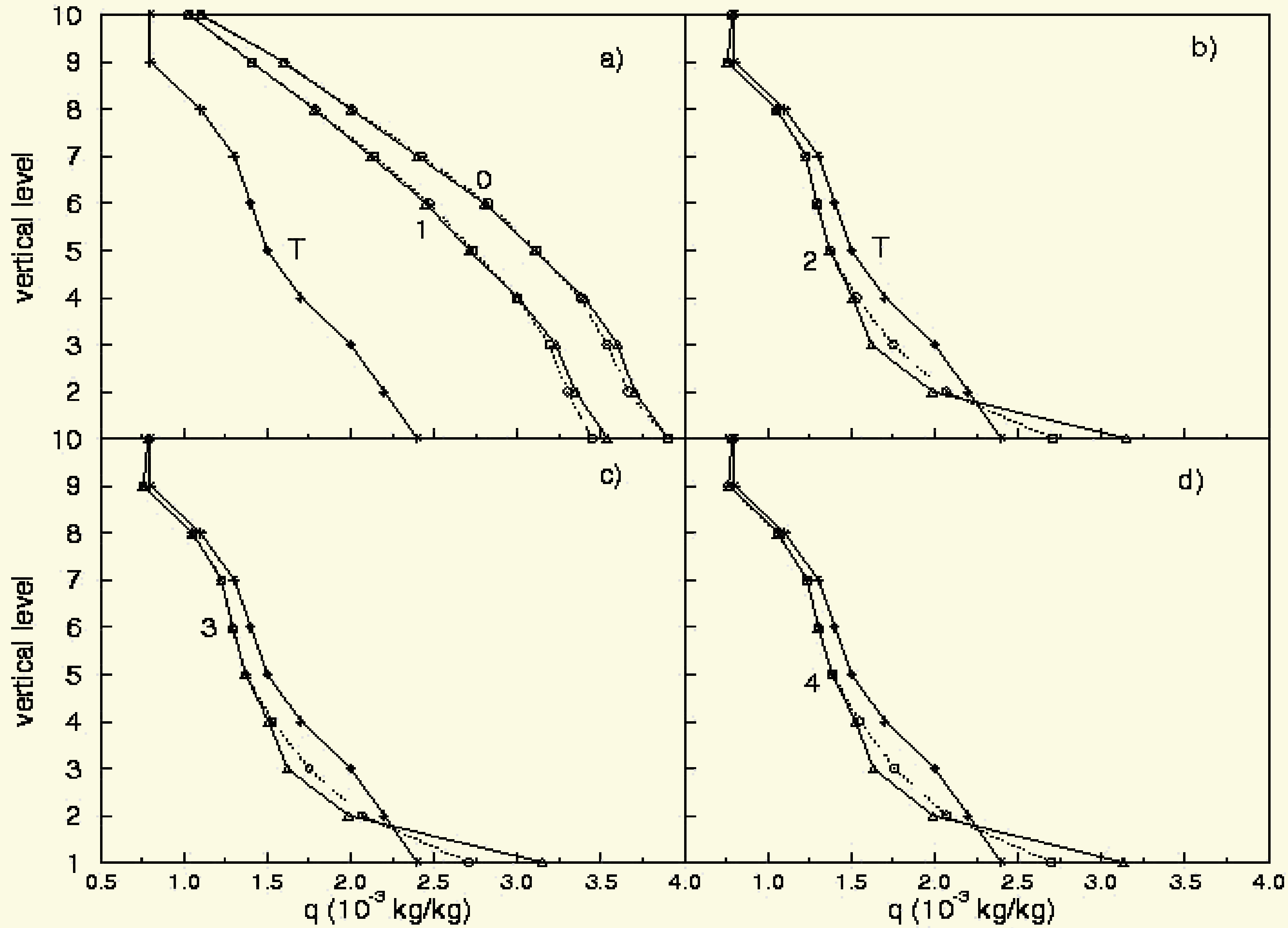
Tracing the route of minimization with L-BFGS for temperature profile



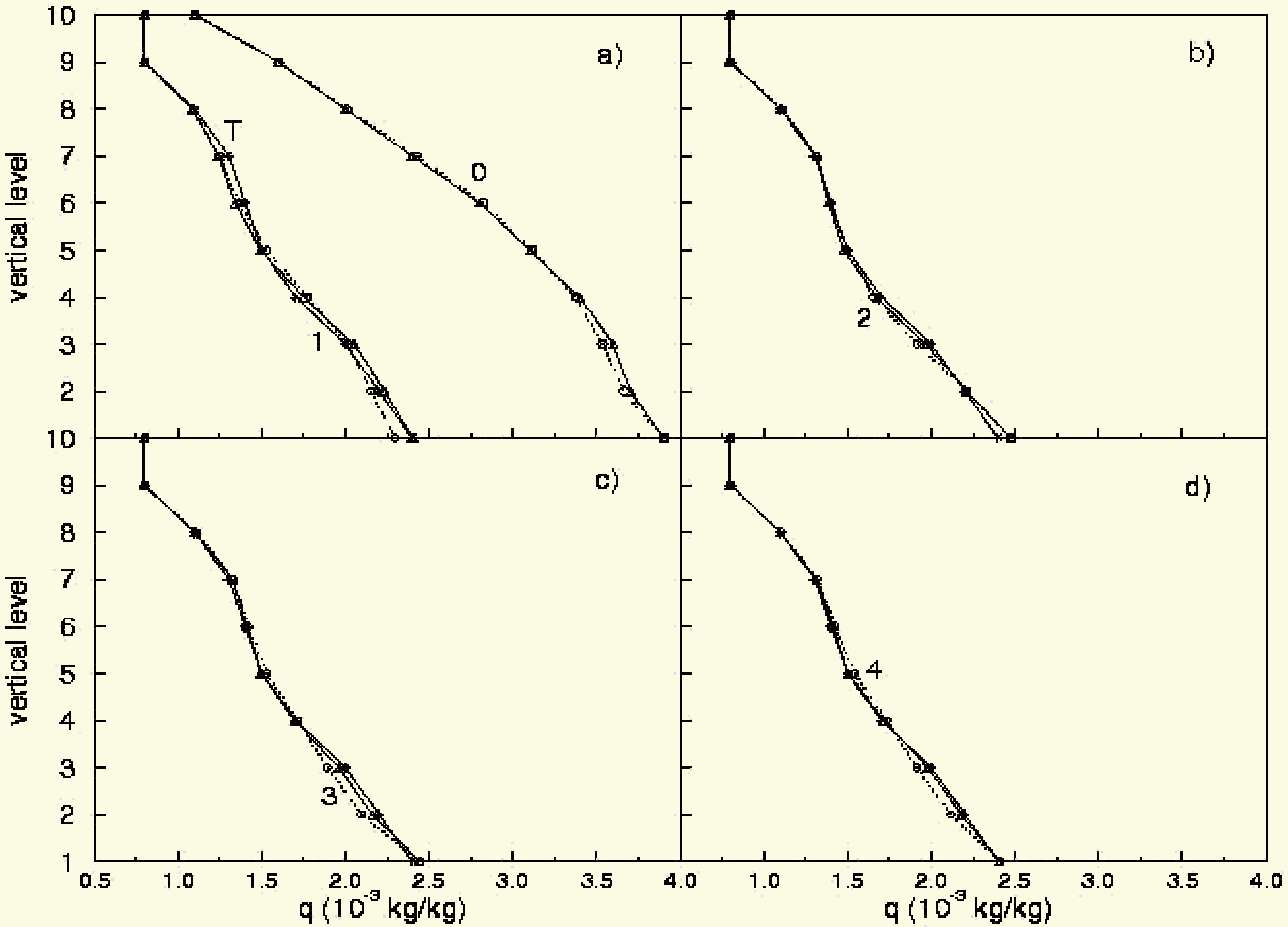
Tracing the route of minimization with bundle method for temperature profile



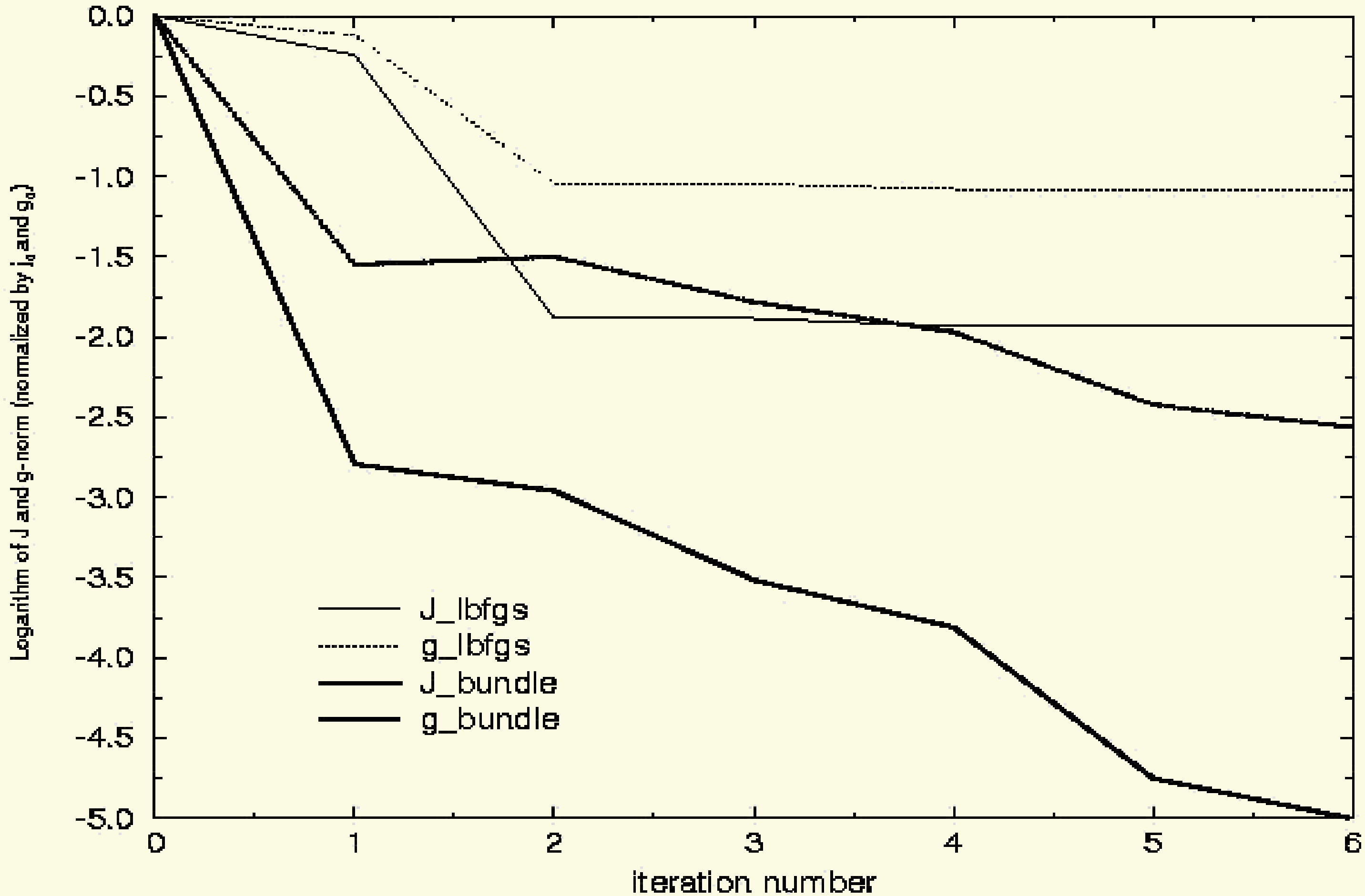
Tracing the route of minimization with L-BFGS for q profile



Tracing the route of minimization with bundle method for q profile

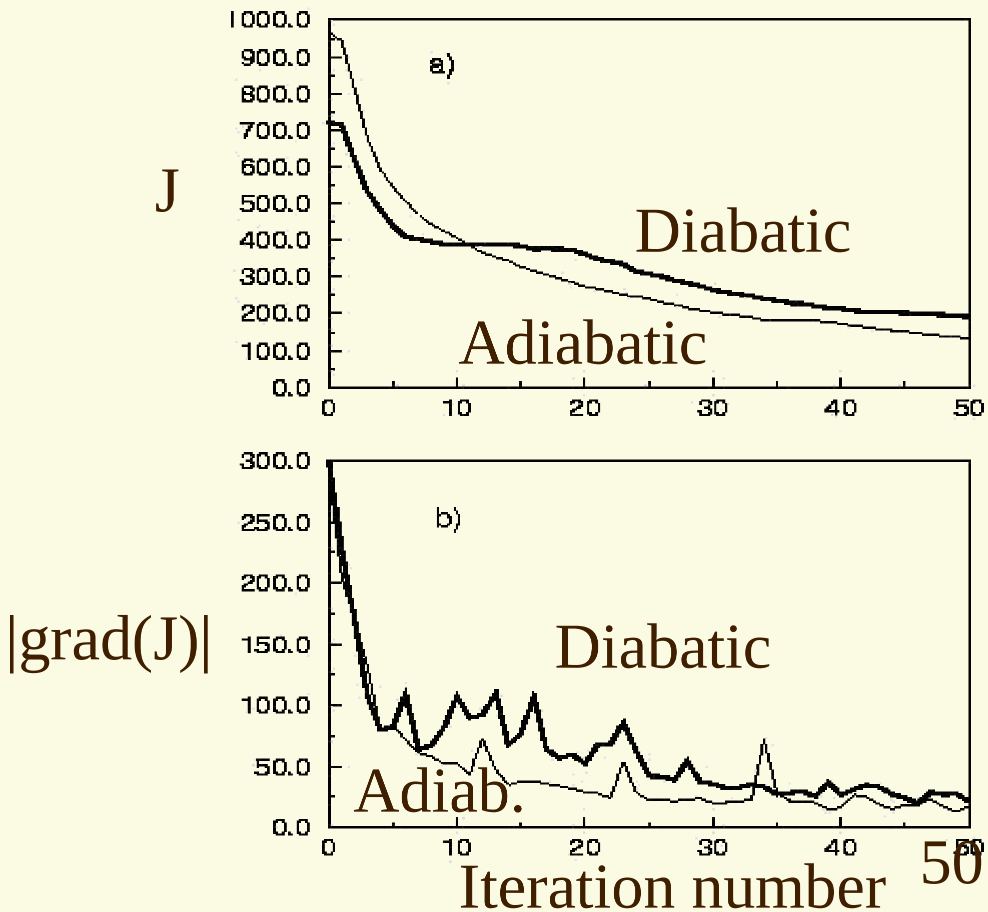


Decreases of J & ∇J for L-BFGS and bundle method

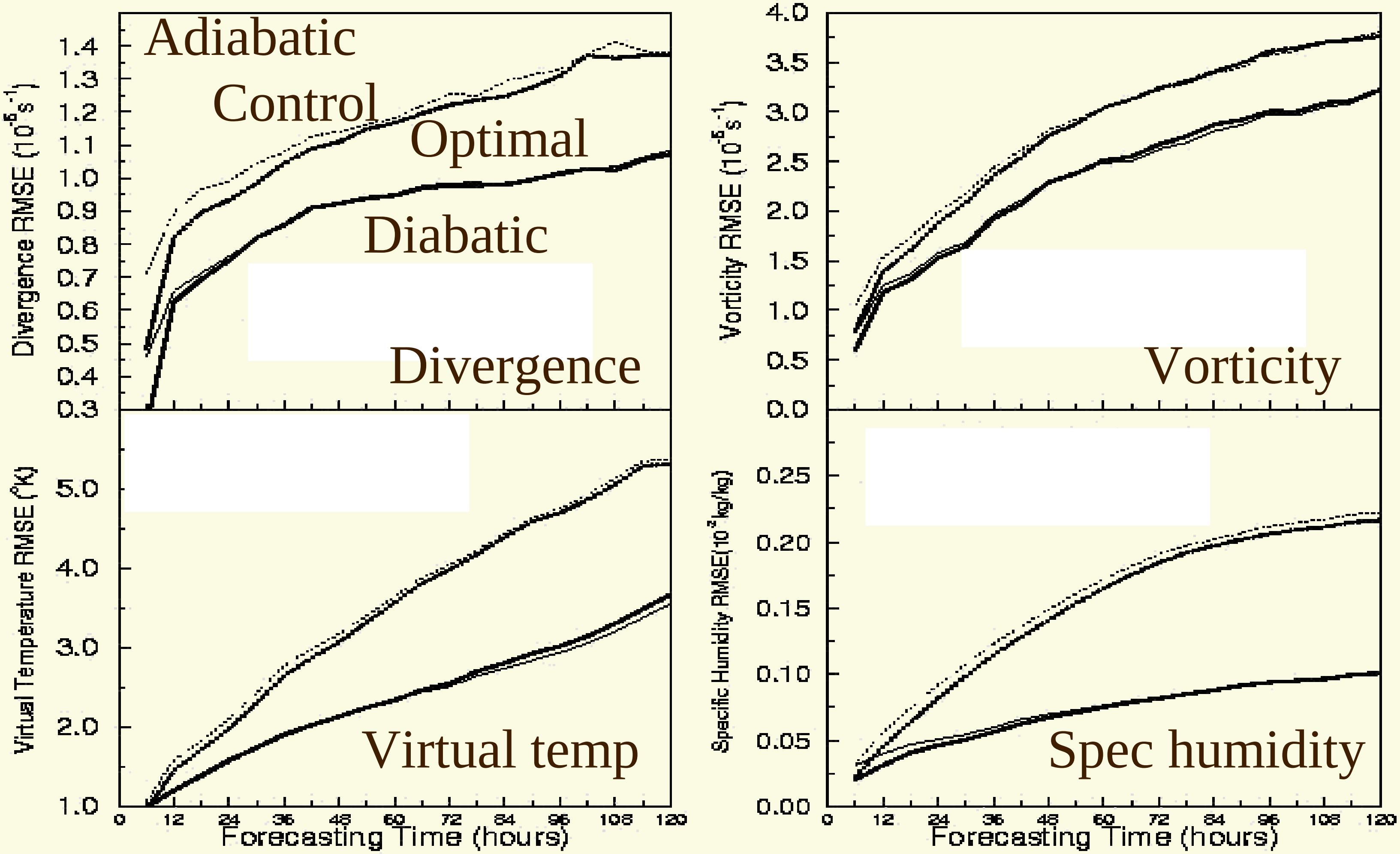


Performance of L-BFGS algorithm with ICs for NCEP global spectral model

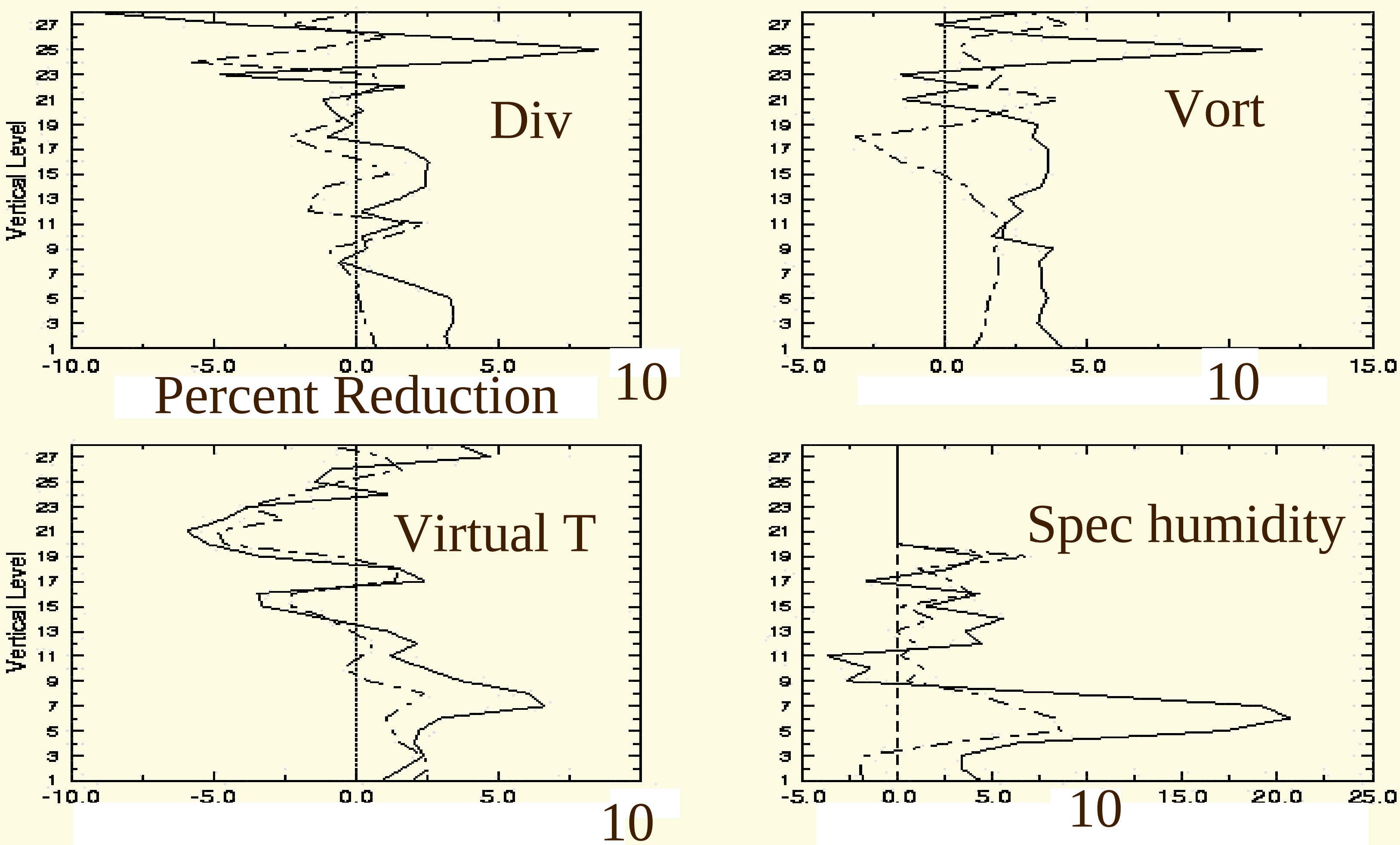
- ✓ Both discontinuity and nonlinearity introduced by parameterized physics affect decrease rate of J



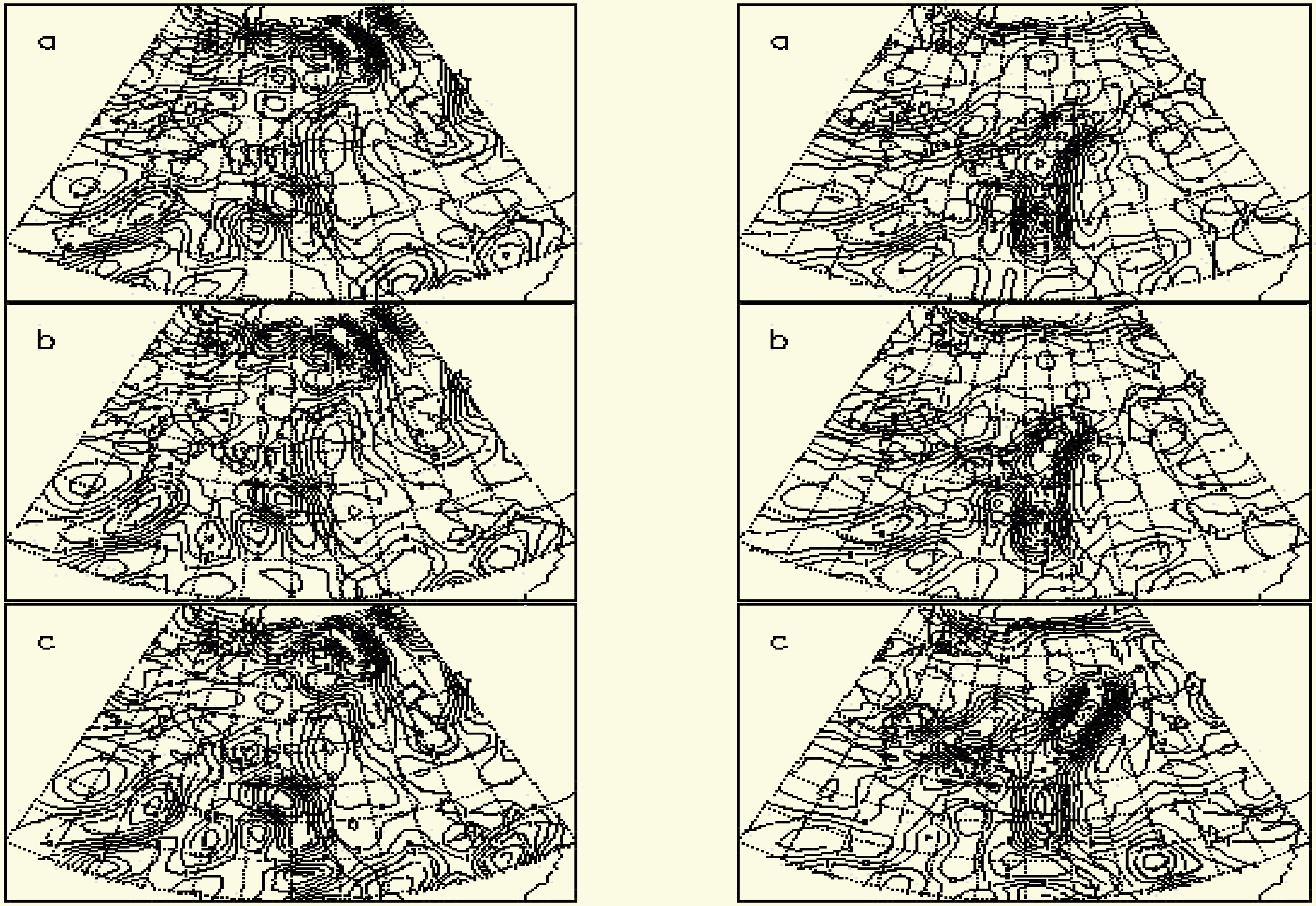
Change of RMSE with forecasting time (out to 5 days)



Vertical distribution of RMSE reduction at 24-h (solid line) & 48-h (dashed line)



Vorticity distribution at sigma=0.8838 over (0E,20N) to (60E,60N) at 6h & 30h forecasts



Summary: Can adjoint correctly evaluate grad(J) when physics are discontinuous?

- Cost function, J, of parameterized physics is piecewise differentiable. Max number of differentiable pieces is $k \cdot 2^n$ for k thresholds and n-step integration, so J becomes rough very fast with more thresholds and time steps.
- Perturbation analysis approach is invalid when a perturbation crosses a discontinuity.
- Adjoint integration is an implementation of the chain rule for differentiation of a complex model, which correctly evaluates gradients (or one-sided gradients) of a piecewise differentiable J.

Summary: Can Newton's method minimize discontinuous cost functions?

- ✓ L-BFGS method (Newton variant) often works well to minimize J, but stationary point may not be global minimum, and even sometimes fails.
- ✓ Bundle method better but twice as slow.

About 4D VAR:

- ✓ Optimal parameter values found by 4D VAR reduce forecast errors only out to 3 days.
  - Imperfect models: affect optimality of ICs and parameters for forecasts beyond optimization interval.
  - Uncertainty: intrinsic loss of predictability with increasing fcst leading time, particularly at small-scales.

Future Work: Classical 4D VAR

- ✓ Evaluate new physical parameterizations by checking cost function and its sensitivity.
- ✓ Test bundle method to minimize cost function for entire model.

My Future Work: Data assimilation for ensemble forecasting (Anderson '99)

- ✓ Given a set of observations, a Monte Carlo implementation of fully non-linear filter solves for a probability distribution of ICs, instead of seeking a single 'best' estimate of ICs.
- ✓ Extending the application to realistic model promises to enhance significantly the quality of ensemble forecasts over a range of spatial and temporal scales.
- ✓ Many obstacles need to be surmounted for the extension.

Thanks to

- ✓ Dr. Jon Ahlquist for his continuous scientific inspiration and many things more than science.
- ✓ My academic committee members: Drs. Barcilon, Navon, Pfeffer and Zou for their generous discussions and advice.
- ✓ Drs. Sela and Kalnay for their persistent encouragement and support.
- ✓ My wife and daughter for hope and love.
- ✓ My friends: Yin, Wei, K-Sris, Zhan, …
- ✓ All who gave me help and encouragement at FSU and NCEP

Questions?